RESISTANT.AI

Resistant Documents - Technical Brief

High-Level Description

Our PDF/Image Forgery Detection system protects business processes and automated workflows from forgery-based frauds. The Resistant AI solution inspects financial documents, bank statements, and other documents submitted to our customers, for signs of manipulation and raises an alert when a modification, file corruption, inconsistency with past behavior, or other anomaly is found by the system.

The system receives PDF documents or images via a secure API, secure Web interface or via email and returns an estimate that categorises the document into one of the 4 levels of trust:

- **Trusted** (the best). Trusted files contain significant evidence that they are authentic (issued by a known and trusted organisation) and unmodified since their issuance. Such files are safe for fully automated processing. To be rated as trusted, the files typically need to come from a known issuer through a fully digital process. Electronic signature is helpful, but neither necessary nor sufficient.
- **Normal** files don't show any sign of integrity tampering, but their origin and authenticity can not be ascertained based on the available evidence.
- **Warning**. The files with this verdict show evidence of modifications or tampering, but the intent of the tampering does not unambiguously indicate a fraudulent intent. This verdict is typical for detections on unstructured or unknown documents, scanned documents or lower-quality documents. The low-quality documents that are blurry, out-of-focus or show flash reflection can be highlighted with Warning as well.
- **High-Risk** (the worst). High-Risk documents contain such signs of tampering that the intentional fraud is the most likely explanation.

Explainability. All verdicts are explained. **Indicators** included in the system response, together with the verdict, guide the investigators in their verification steps, and most of them point-out the locations of specific changes if possible. This significantly reduces the investigation time, allows for the quick resolution of false alarms, and provides independently verifiable claims for possible formal investigation.

API and Connectivity

The primary method of service delivery is the Software-as-a-Service model, based on the public cloud and fully hosted and managed within the Resistant AI AWS account. Alternative methods (ranging from third-party provider, customer cloud accounts and on-premise deployments) are available upon request and are represented in a separate document.

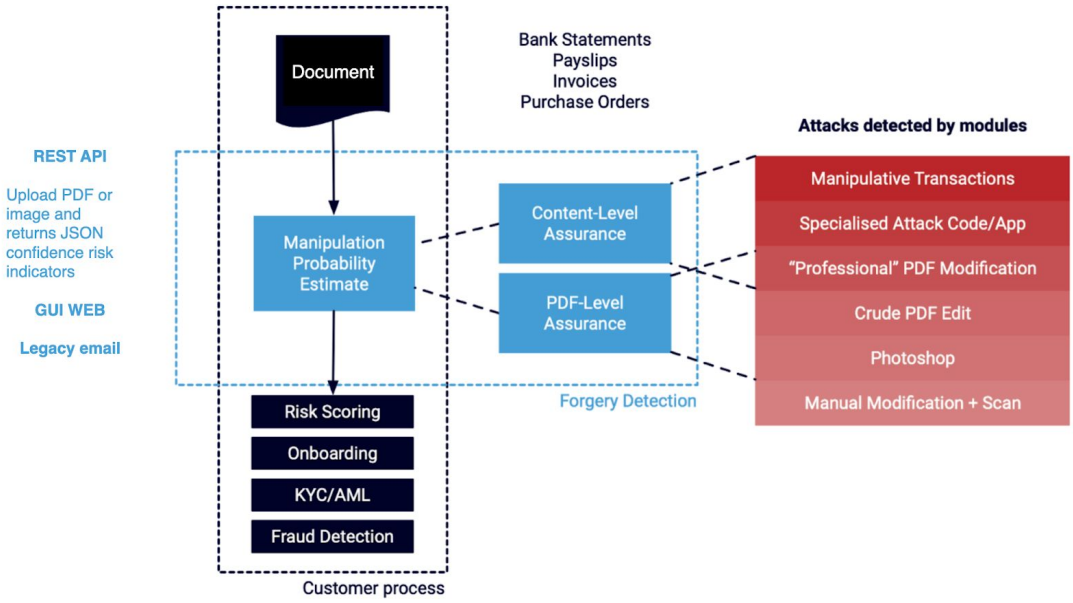
The Resistant Documents service uses a very simple **REST API** that allows the asynchronous submission of files and reception of analysis results. The *API documentation* is provided in a separate document.

A **GUI application** is available for manual use of the service, and for use by the fraud investigation staff. As a matter of principle, all of the information, visible in the GUI, is included in the result sent by the API, and available in JSON format. This decision is based on two principles. First, we want to make the whole verdict, including the indicators presented to the investigator, to be recorded for future reference. Second, we want the GUI to be stateless and fully contained in the browser session, in order to prevent any

accidental leakage of information. Following the same principle, our system never receives any record of investigator activities within the browser, with the exception of the initial request to obtain the verdict and the indicators.

In order to facilitate **batch-mode usage** of the system and processing of large archives, we support direct integrations with document archives on the customer side by supplying customised integration stubs.

Email integration, where documents are submitted via email and/or the response to previous uploads is returned via email, is also supported to connect to legacy systems. However, this option is not recommended as it does not offer sufficient confidentiality guarantees for the data in transfer.

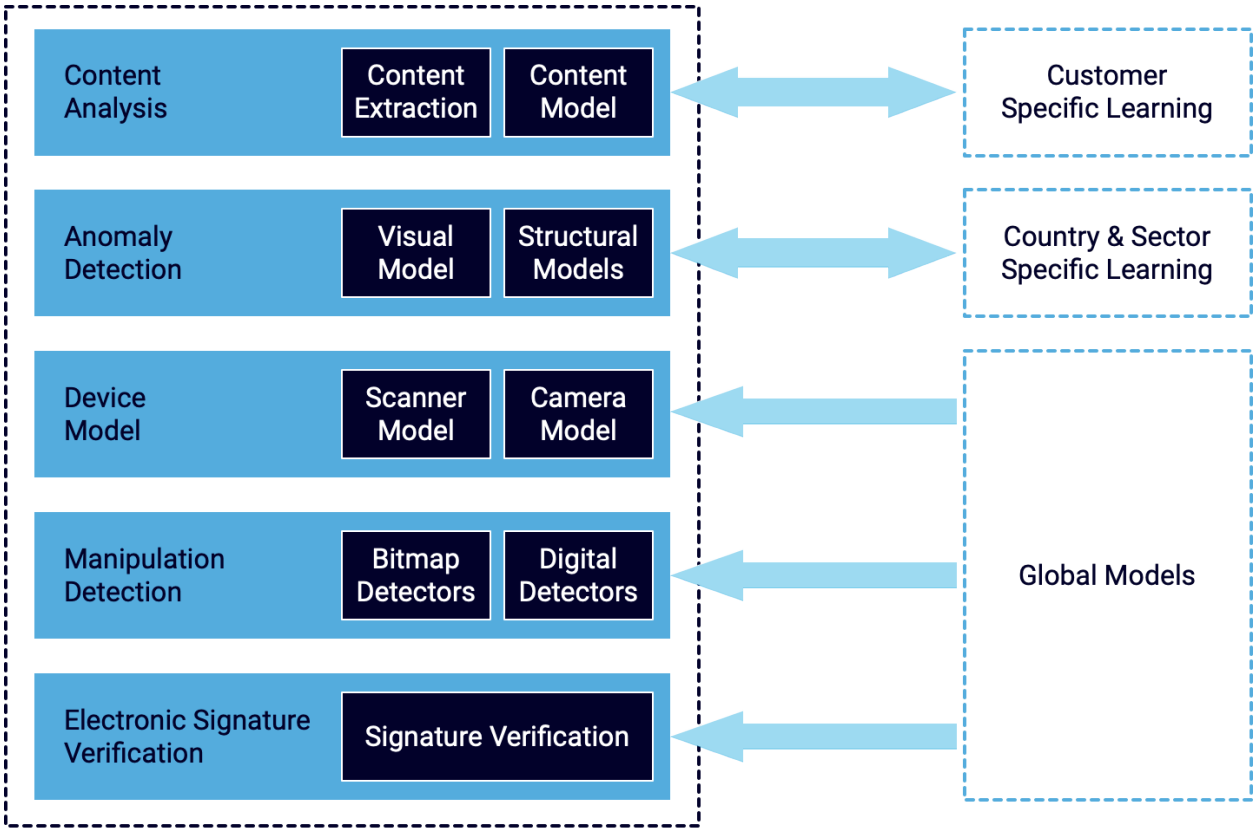


System Architecture and Functional Modules

The key engineering problem with forgery detection is the diversity of documents, their forms (e.g. digital originals, scanned or photographed), and even greater diversity of attack and fraud techniques and goals. In order to cope with this complexity, without requiring an excessive number of training samples, we break the problem down into manageable sub-problems. Then, we build specialised decision modules that deal with these problems.

The impact of such a breakdown should not be understated. It allow us to:

- make the training process faster and more robust,
- achieve much better **initial performance** (out-of-the-box) for new deployments,
- ensure that the training process of the majority of the system **does not rely on any GDPR-protected information** or other sensitive information, and
- ensure that the results are explainable and well justified, despite being produced by advanced ML techniques.



The Resistant Documents system is structured into 5 layers. Each layer is composed of several logical modules. Starting from the bottom, we perform the following types of verifications:

- **Electronic Signature Verification.** This module checks the electronic signatures on the PDF document, or its parts, and verifies the signature and the signatory against the past practice of the same source and authoritative services. Separate detectors have been designed to discover tampering with the signed text after the rendering and highlight partially signed documents.
- **Manipulation Detection**. This set of detectors is designed to estimate the risk of modification of the PDF file or image regardless of its content or origin. The estimate of the risk is based on the identification of modification artifacts, traces of editing, and structural anomalies apparent on a

single-file level. This type of detection can be performed globally, using the experience with software editors, formats, and other artefacts directly related to the manipulation.

- **Device Model**. This layer models the acquisition of the original document (in paper form) into digital form on different models of scanners and mobile phone cameras. It has been designed to identify the files that have not been modified since their digitisation. This allows us to provide more confident Trusted and Normal verdicts regarding the photographs or scanned documents.
- **Anomaly Detection**. This feature module adds detectors that increase the likelihood of detection of sophisticated fraud and insider threats. The documents analysed by the system are categorized and stochastically attributed to sources. Each document source (such as a bank, payroll company, accounting software, or a specific business process of a partner company) is modeled and every new document attributed to this source is checked against the expectations embodied in the document model. Therefore, we can discover additional (professional) categories of fraud, where the document has been produced without explicit modification indicators, but differs from the past practice. This module is also the core of our ability to deliver **Trusted** verdicts mentioned above.

Overall, the system contains more than **100 detectors** across the modules presented above. Due to their number, we are not able to provide a detailed explanation for each of them, but we would rather convey the main principles by covering the methods used in their development and then by presenting a specific detection use-case.

Machine Learning Techniques

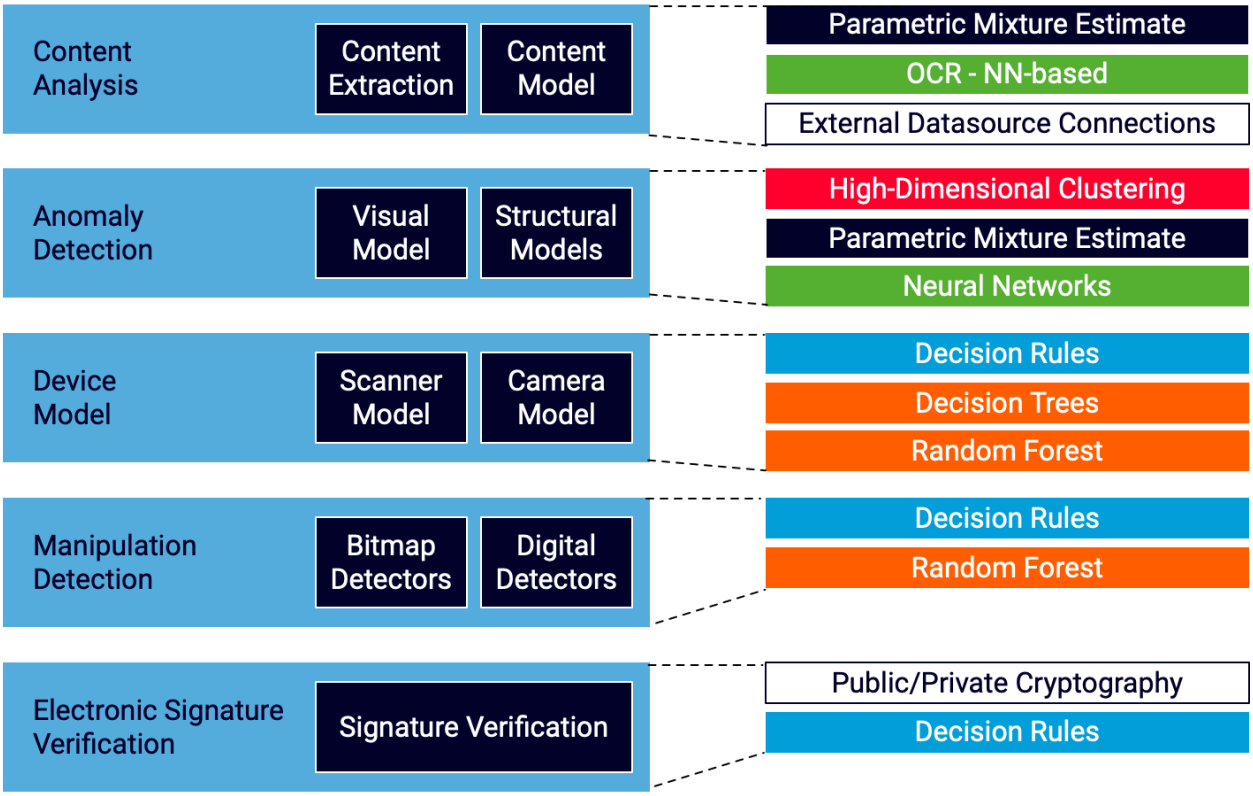
Our system solves difficult decision problems in an inherently adversarial setting. In doing so, it combines different classes of machine learning algorithms in order to reach reliable conclusions. Besides the obvious method differences (unsupervised clustering vs. CNN vs. random forests), the main and actually far more important difference is the scope and time characteristics of the training of individual models. The methods used in the **Manipulation detection** module are typically pre-trained and applied uniformly across all documents from the whole customer base. In this module, we aim for simplicity, stability and robustness, and most of the research effort is invested into feature selection and robust pre-processing.
More complex ML techniques are used for result aggregation and verdict generation and rely primarily on the results of lower-level classifiers, with secondary use of original input features for decision context discovery.

The **Device Model and Anomaly Detection** layers have been intentionally designed in a complementary way. They use similar techniques to address different aspects of document lifecycle and different document types, i.e. digital originals vs. scanned documents.

Anomaly Detection combines High-Dimensional clustering with Neural Networks to attribute the documents to a specific software, category of document and the source - most often the company issuing the document. We intentionally use different techniques on different data here. The software attribution component has been designed as benevolent, in order to maximise attribution to software based on visual similarity - it works for scanned and digital documents alike. The source/company attribution intentionally relies on the content of the document identifying the issuing company and has been designed to be as close to human understanding of the information as possible. Once we attribute the document to a specific source and document class, we use a high number of statistical anomaly detectors in order to detect outlying, novel or anomalous characteristics of the individual documents.

Device Model is applicable to scanned and photographed documents. It uses relatively simple machine learning techniques (Random forests and decision trees) on a very large scale, in order to identify the typical scanners, mobile phone cameras and other devices used to digitise a paper document. The model then builds a two layer model analogous to the issuer-based document Anomaly Detection. The point of the first layer is to identify the *apparent* device of origin, i.e. the scanner that the document claims it originates from. The second, much more detailed and restrictive model then compares the document to

similar documents acquired by the scanners of the same model, manufacturer or OEM sensor provider and verifies whether the properties of the image fall into the expected sets of values.



Selected machine learning techniques as applied in different modules. Only key techniques are shown.

The methods used for **Electronic Signature Verification** do not directly rely on machine learning approaches, but re-use the results of document attribution to a class and customer provided by the Property Identification and Anomaly Detection layers. The verification itself is a straightforward application of cryptographic techniques and diligent verification of document properties in order to verify possible manipulative elements covering the signed text.

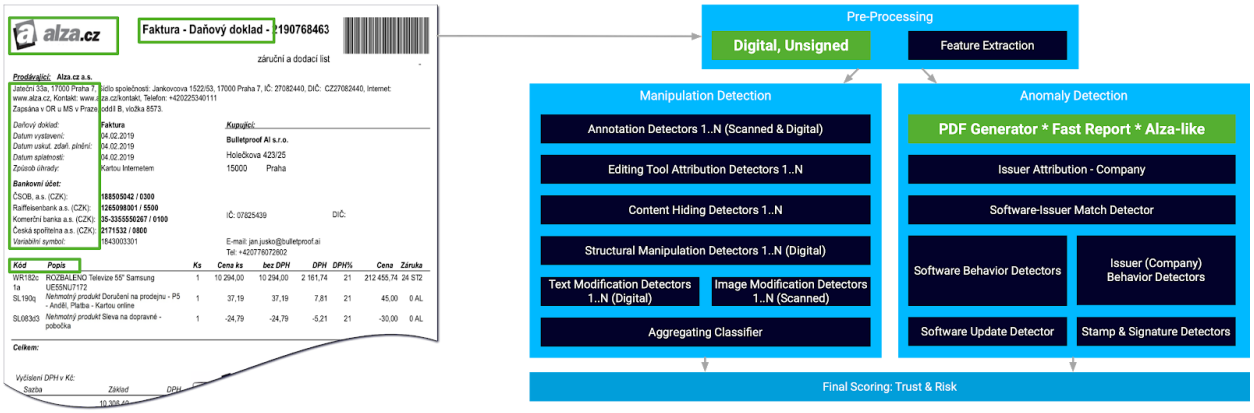
Our approach to machine learning does not rely on a single silver-bullet approach. We decompose the problem, identify the simplest and the most robust techniques for each task, and build the system as a layered ensemble of classifiers working together, as we will see in the example in the next Section.

Example of Detection - Digital PDF Document

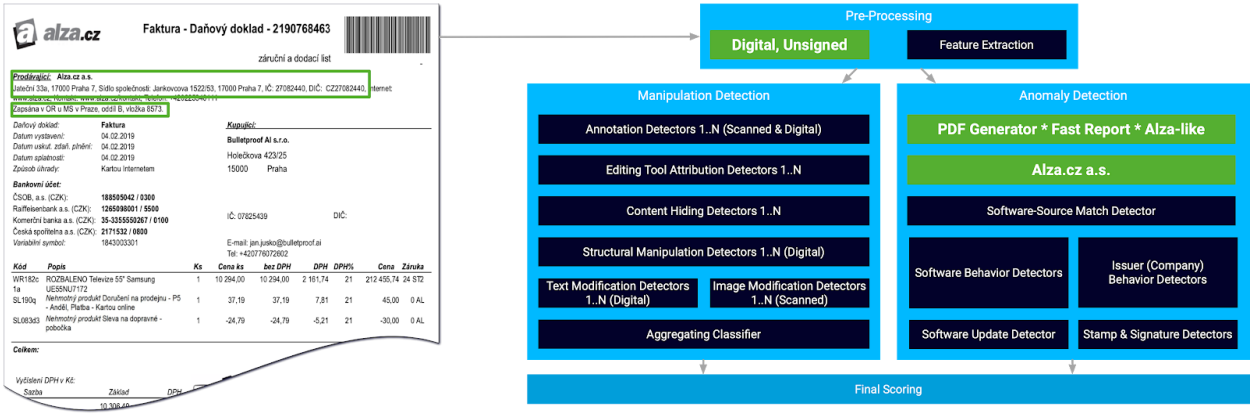
In this section, we will walk through a single instance of fraud detection by our system. The scenario covers a digital modification of a digital PDF invoice, produced and emailed by a specific supplier. This kind of fraud is typical in factoring (PO modification), insurance fraud (invoice/bill modification), SME lending and car financing (invoice modification).

Please note that the below diagram only involves 2 out of 5 system layers: Manipulation Detection and Anomaly Detection. The other layers would be used on other documents, with different characteristics.

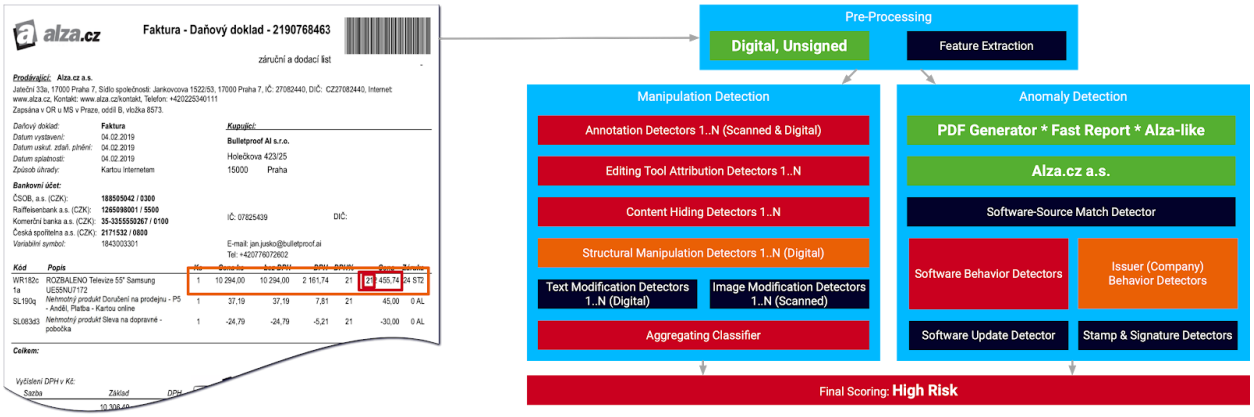
Once the document is identified as Digital PDF, Feature Extraction extracts the information from the document and represents it in a form suitable for detailed analysis. We extract several thousand features from the document and pass them to the Anomaly Detection and Manipulation detection modules.
The first layer of the Anomaly detector assigns the document to a specific class using its appearance and the origin. We have finished the attribution by now: we have identified the software information (from structure and meta-data), as well as document cluster membership.



Once we know the attribution of the document to one or more classes, we make the next step and attribute the document to a specific source (company). For each cluster of documents, we use a cluster-specific approach to select the information elements in the document that identify the source, such as VAT number, company name, address or a phone number.



Then, we use the attribution information to select the right anomaly detectors to apply to the document and we also apply the manipulation detection techniques at the same time. As the document has been modified, the results begin to appear.



The first check actually works out - the document original has been produced by the software system previously associated with the source and document type. Then, the anomalies start to surface. The document contains annotations - a first for a document from this software and this source. The document also contains fonts not consistent with document origin, besides the other indicators.
In the manipulation detection, the (independent) detectors trigger as well - presence of annotations/text annotations is noted. This could still be fine, but their adjacence to the original text triggers content hiding detectors. The type of the editor used to perform the modification is identified and reported, both thanks to the metadata, as well as the structural changes in the document itself. We have therefore scored the document as altered, reported the modification software and highlighted the change in the document both graphically and in the structured output.

Please note that the above example only shows a fraction of the techniques available for detection of forgeries on digital documents. Upon your request, we can provide more in-depth information about the other detectors.

Image Documents Modification Detectors

The analysis of photographed or scanned documents in JPEG, PDF, PNG or other formats is significantly different from the analysis of digital documents outlined in the previous section. We need to consider not only the document itself, but also the properties of the device used to digitise the document, its interaction with the document and possible interactions with "alternative" (fraudulent) documents with the same superficial look.

This interaction of multiple transformation steps is what makes this task much harder than the detection of modifications of purely digital files. The fraudsters can (and do) engineer the subsequent transformations specifically to make the discovery of modifications very difficult, by artificially lowering the resolution and increasing compression ratios. This is why detailed device modelling capability is so critical. Knowing that some transformation is very unusual for a specific image source can provide very valuable hints regarding the existence and the intent of the fraudster.

Manipulation Detection

Manipulation detection of photographs and scanned documents breaks into two distinct sub-fields: Detection of manipulation before the digitisation, and detection of manipulations of the already acquired digital image in the computer. In this section, we will address the latter.

More than 99% of scanned or photographed documents are directly or indirectly (scanned PDF files mostly contain JPEG files in a PDF container) in the JPEG format. This allows us to design a whole range of methods designed to detect the manipulation of the JPEG images using the knowledge of the compression mechanism used in the JPEG format. This allows us to develop numerous detectors:

- **Error-Level Analysis**: A very common method that detects the noise pattern change caused by localised edit operation in the JPEG image. We use the method as a complementary detector to the more advanced methods below.
- **Localised Discrete Cosine Transformation Coefficient Analysis**: We detect image elements that have richer spectrum representation than the corresponding unmodified elements of the image.
- **Noise boundary analysis**: This method discovers the artificial edges in the image that would be unlikely to appear naturally, but frequently appear as a consequence of image modification.
- **Copy-Move Detection**: This method inspects all the characters within the document and detects the cases where a specific character is reused. We use the similarity of digitization and thermal noise patterns to confirm that it has been copy-pasted from another part of the document.
- **Detection of multiple compressions steps**: This detector detects the missing density peaks on the image histogram that appear as a consequence of the successive application of the compression step.
- **Detection of editor and editing artefacts**: These detectors have been designed to look for artefacts within the PDF or JPEG mage caused by the image editor. The artefacts can be of multiple forms - some editors leave meta-data, other software leaves (very helpfully) the copy of the original content to allow undo operation, while the others transform the data with distinctive mathematical operation and coefficient combinations. All of these types of indicators are detectable by our system.

Device Modelling

Device modelling is a necessary step that allows us to discover very advanced attacks, while not increasing the false alarm rate on the normal noisy, yet legitimate inputs. In order to achieve that, we model all common scanners and phones, used by our customers (and their users). We process a wide range of documents to extract the typical image characteristics associated with each device, such as resolution, compression, white-balance compensation, image (de)composition into segments, compression levels and other parameters that are reflected in the image files. This anomaly detection-based learning process then builds expected behaviour representation.

At runtime, it analyses the images submitted to the model and verifies whether the inputs that claim to originate from a specific device closely match at least one of the known behaviour profiles built for the device. Really close match covering a sufficiently large parameter space of a well-known device can lead to (slightly more narrow) Trusted verdicts that confirm that the image has not been altered digitally, between the acquisition and the submission for processing.

Pre-Digitalisation Manipulation Detection

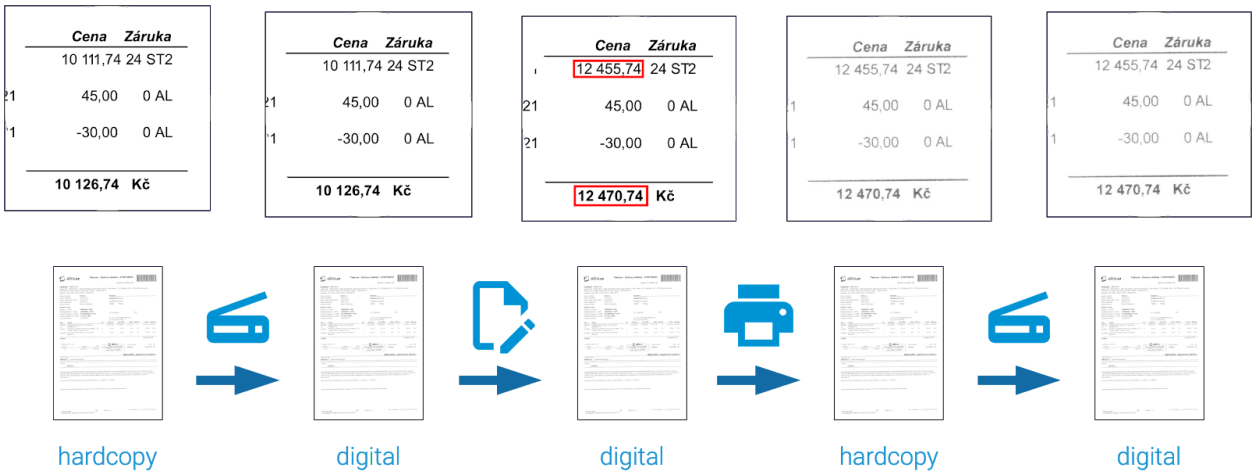
Document modifications performed prior to scanning or photographing the document is the most difficult category to detect. This is due to the fact that the fraudsters would intentionally use a digitisation technique that would deliberately restrict the amount of the information available to the system, such as

low-resolution, low-contrast, bad lighting conditions or low-quality settings of the image acquisition device. On the other hand, the task is not impossible and the Resistant Documents system contains several methods that verify whether the document is consistent with the past practice of the issuer and whether it contains effects typically associated with tampering. These detectors include visual detectors, such as font similarity detection, line and text block alignment detection and other methods that assess the visual conformity of the document.

But what would happen if the document were to be scanned, modified digitally and then scanned again? This is exactly the question we are addressing with "print-scan-print" or "double print" detector, that works as follows.

The idea behind the print-scan-print detector is very simple. We start with a (very pessimistic, yet reasonable from a secure design perspective) assumption that the change of the document is in itself undetectable by all of the other detectors mentioned above. This leaves with two options how to catch the fraud. We can either detect the fraud workflow, or we can detect the content change.

Print-Scan-Print detects the workflow. It uses spectral analysis to discover the image artefacts related to document re-printing after the (digital) change. By modeling the printer behavior, when working with scanned content, and its combination with the subsequent scanning, we can discover the characteristic frequency related to halftone representation of the re-printed document, as we show in the picture below.



Content Extraction and Analysis

Content Extraction

Resistant Documents is able to extract textual information from the analysed document and either:

- return the content using the API call,
- use it for content-related security validation within the system, or
- use the information for verification in a 3rd party system, such as government registers, blacklists or scoring agencies.

We don't provide a generic OCR solution for unspecified documents, but rather provide a solution designed specifically for security and fraud analytics[1]. In this solution, we extract very specific information from an authentic and unmodified document and further validate it before returning the information and/or using it in content verification.
As our first step, we leverage the document class identification that we use in the Anomaly Detection validation. Class identification works equally well for digital documents, scanned documents or photographs. It identifies the class of the analysed document and we use the class information in order to select the target regions to acquire.

Then, we extract the text from the target regions and represent them in structured or unstructured way. The extraction technology depends on the document format.

- For digital documents, we parse the text content from the document itself in order to prevent any OCR errors. This results in near-100% precision (most of the extraction failures are related to fraudulent document modifications) and high speed of extraction.
- For photographs or scanned documents, we use the OCR engine embedded in our solution to extract the information. Our approach has been optimised for robustness to manipulation and uses several successive runs of the algorithm to transcribe the content as accurately as possible. In case of doubt or unstable output, the engine is set-up to reject the document and abort the transcription.

The extracted information is available in the JSON format. It is returned in a separate API endpoint, so that it can be segmented away from the fraud detection result. The reason is that the fraud detection result does not normally contain any PII information according to the GDPR documentation, while the content extraction endpoint contains nothing but such information.

Content Analysis

The content that we extract from the analysed document can be used for additional validation checks. These chacks are of three kinds:

- Internal checks performed by detectors inside the Resistant Ai solutions, such as IBAN verification on supplier invoices against the past practice, or verification of checksums and balances within the single document.
- External checks performed by Resistant AI against third-party sources, such as government databases, scoring agencies, rating agencies. VIN verification is the canonical example of this.
- External checks performed by partners of Resistant AI against the third party sources. Partner checks allow the use of customer-side databases and additional private information to take reach the optimal decision.

All of the above methods are typically customised for a given vertical and geography combination.

New Customer Experience

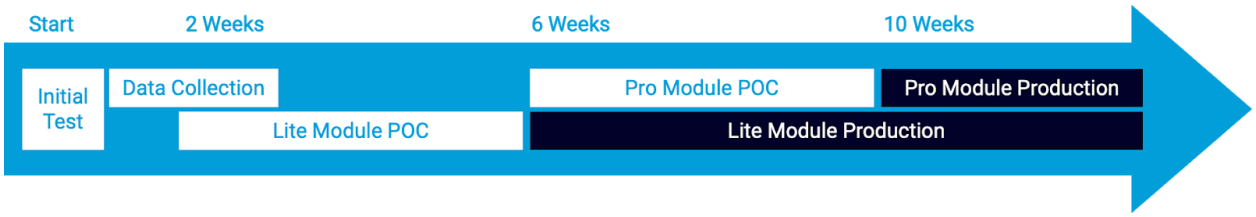
The system has been designed as a self-improving and constantly learning system behind a constant, non-changing API. It provides significant customer value from the first day of customer deployment. Manipulation detection works on the nominal level from the beginning. This is also the case for the Signature Validation module. The system is able to detect standard attacks and can therefore detect a vast majority of external, non-professional attacks. The Trusted verdicts are solely based on electronic

[1] In case of need, we can integrate with OCR solutions available as open-source (Tesseract) or commercially, where we can leverage our partnerships with UiPath or BluePrism or use the OCR capabilities offered by the AWS Textract service.

signatures and limited assessment of globally identified Properties of acquisition devices, software, libraries, and system configurations.

As the system processes more and more documents, it builds and improves anomaly detectors for many sources and becomes more precise. The length of this initial period is typically a couple of weeks, depending on the richness of the sources and the rate of arrival of new documents. The Anomaly Detection module kicks in and progressively builds models of expected behaviors for business partners, starting with the most common document issuers. It therefore reduces some false alarms originating from software" "creative" use of document formats and features more typically associated with fraud, as these can become more acceptable and expected for a given partner after sufficient experience. It also starts providing Trusted verdicts based on consistent history of behavior for each source and allows higher level of automation on these documents. On the other hand, it provides tighter security bounds and would consider any deviation from established behavior as a risk factor.

The initial phase can be considerably shortened or completely eliminated by bulk-loading the system with an appropriate volume of existing documents representative of the future inputs.



Deployment in new Territory and Vertical

Many sources (software packages, large banks and their documents fingerprints, ...) are shared between deployments and allow faster bootstrapping. The models of these documents (obviously completely independent of individual customer's or user's private data) can be reused and shorten the learning period of a new deployment. This is the main point of the Property identification sub-module.

When entering a new territory and vertical, the system needs to identify new issuers (document sources) and attribute them properties that may be identified globally. In case of significant use of region-specific or industry-vertical-specific software, the system may need to identify and model new properties and associate them with the source.

Limitations and Restrictions

Our system is not a panacea and has not been designed as a fully automated, human-less solution for fraud detection. It provides an assessment of documents and scores them to estimate the appropriate level of Trust and Risk, so that the document can be assigned either to a fully automated workflow, rejected as a likely fraud, or assigned to a human analyst that would make the decision. The system is meant to enhance and scale-up human decision making, not to replace it completely. The verdicts provided by the system should be considered in the context of transaction risk, counterparty risk and other characteristics only known to the customer, who remains solely responsible for the final decision.

Our service has been designed to detect forgeries and modifications of specific, yet very large categories of documents: forms, bank statements, purchase orders, invoices, payslips, and other documents generated by enterprises or large organizations in the course of their business. The service also covers automatically, software-generated documents issued by smaller organizations. The software may be used on other documents, but has not been designed or tested for other purposes.

One particular point to note is the impact of document origin and type on estimation accuracy. In general, our system provides the best quality of verdicts for digital PDF documents generated by automated processes. Paper documents scanned into PDF form, screen photographs, or other images and bitmaps can be very reliably validated for the modification after they have been digitized and our confidence in these verdicts is very high. The manual modifications performed prior to scanning and digitization are harder to detect, but this is an area of very intense research and we do provide several detectors designed solely to discover this category of fraud.

Privacy Protection

The lower, technical layers of the system have been designed to **exclude the** actual **content** of the document from consideration as much as possible, in order to prevent the storage, processing, and analysis of any information that might be covered by the GDPR regulation. The features used by the system concentrate on software artifacts, graphical elements, page layout, and similar features. The Anomaly detection models are based on document sources - the system models the software stack and processes of large organizations that issue the documents that we verify. These organisations are not subject to specific privacy protections. In the limited case of content analysis for bank statements and transaction documents, we only model the relationships between the pseudonymous entities and emphasize the active entities that can be unambiguously connected to large organizations.